Exploiting Novelty, Coverage and Balance for Topic-Focused Multi-Document Summarization

Xuan Li[1,2], Yi-Dong Shen[1], Liang Du[1,2], Chen-Yan Xiong[1,2]
[1]State Key Laboratory of Computer Science, Institute of Software,
Chinese Academy of Sciences, Beijing 100190, China
[2]Graduated University, Chinese Academy of Sciences, Beijing 100049, China
{lixuan,ydshen,duliang,xiongcy}@ios.ac.cn

ABSTRACT

Novelty, coverage and balance are important requirements in topic-focused summarization, which to a large extent determine the quality of a summary. In this paper, we propose a novel method that incorporates these requirements into a sentence ranking probability model. It differs from the existing methods in that the novelty, coverage and balance requirements are all modeled w.r.t. a given topic, so that summaries are highly relevant to the topic and at the same time comply with topic-aware novelty, coverage and balance. Experimental results on the DUC 2005, 2006 and 2007 benchmark data sets demonstrate the effectiveness of our method.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Abstracting methods*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms

Algorithms, Experimentation, Performance

Keywords

topic-focused summarization, topic-aware novelty, topic-aware coverage and balance, relevance measure

1. INTRODUCTION

Automatic document summarization has received great attention in recent years, due to the explosive growth of documents on the World Wide Web. One specific task of document summarization is topic-focused summarization, which is to generate summaries that are relevant to a given topic.

A good summary should satisfy the following three requirements: *novelty* ([1], also called diversity in [4]), *coverage* [6, 4] and *balance* [4]. They require that sentences in a summary should describe different contents, cover every important aspect and pay more attention to more important aspects of the documents, respectively. When a topic is introduced for topic-focused summarization, we believe that *topic-aware novelty, coverage and balance* are preferred. These requirements are defined as follows. Topic-aware novelty requires that sentences in the summary should overlap less in topic-related contents. Topic-aware coverage requires that the summary should cover different aspects of topic-related contents. Topic-aware balance requires that different aspects of topic-related contents should have the same relative importance in the summary as in the documents.

In this paper,we propose an unsupervised topic-focused multi-document summarization method. It differs from the existing methods in that the novelty, coverage and balance requirements are all modeled w.r.t. a given topic, so that summaries are obtained from sentences that are highly relevant to the topic and at the same time comply with topic-aware novelty, coverage and balance. In the method, the relevance of a sentence to a topic is modeled as a topic relevance probability that takes into account the novelty requirement, while coverage and balance are modeled by a sentence selection preference function. Our method is quite effective. Extensive experiments on the DUC[1] 2005 – 2007 benchmark data sets demonstrate that our method outperforms representative existing approaches in all specified evaluation measures.

2. RELATED WORK

We briefly mention some related summarization methods, especially the those that considers novelty, coverage and balance requirements.

Li et al. [4] treats summarization as a supervised sentence ranking problem, where structural SVM is applied. Novelty, coverage and balance requirements are incorporated into the sentence ranking process. However, [4] focuses on generic summarization, with no topic-related information considered. Unsupervised methods such as TextRank [5], LexPagePank [3] and ManifoldRank [8] rank sentences based on graph ranking algorithms. [8] uses the maximum marginal relevance (MMR) [1] like method to penalize overlapping of sentences, where all words are treated equally whether or not they are related to the topic. The method proposed in [9] is closely related to our work in that it achieves topic-aware novelty. Similarity between sentences are biased toward the topic using the proposed query-sensitive similarity in the paper. Thus only overlapping of the topic-related contents is

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.
Copyright 2010 ACM 978-1-4503-0099-5/10/10 ...$10.00.

[1] `http://www-nlpir.nist.gov/projects/duc/index.html`

penalized if the MMR method is applied. Some summarization methods such as ClusterCMRW, ClusterHITS [7] and [6, 10] utilize clustering techniques. The rationale behind the clustering-based methods is that different clusters represent different aspects of the documents, which should all be covered if possible.

3. A UNIFIED FRAMEWORK FOR TOPIC-FOCUSED SUMMARIZATION

In this section, we introduce a sentence ranking method that incorporates topic-aware novelty, coverage and balance into a unified framework.

3.1 Topic Relevance Probability

Topic-focused multi-document summarization tasks are often viewed as sentence ranking problems. If we can estimate the topic relevance probabilities of the sentences, we can rank sentences according to their relevance probabilities. We denote the relevance probability of a sentence s to a topic t by $P(R_{t,s} = 1|t, s)$, where $R_{t,s} = 1$ denotes that t and s are relevant and $R_{t,s} = 0$ otherwise.

In this section, we propose a method for estimating $P(R_{t,s} = 1|t, s)$. We model the *information need* of a topic as a set of topic words $t = \{w_1, w_2, ..., w_m\}$. For convenience, in the following by a topic we refer to its information need. A sentence s is considered to be relevant to a topic t if s covers some information need of t, i.e. some $w_i, 1 \leq i \leq m$ in t. We denote by $R_{w,s} = 1$ that s cover w and $R_{w,s} = 0$ otherwise. Therefore, the topic relevance probability $P(R_{t,s} = 1|t, s)$ is defined as:

$$P(R_{t,s} = 1|t, s) = P(\exists w_i \in t \text{ such that } R_{w_i,s} = 1). \quad (1)$$

We assume that $R_{w_i,s}$ and $R_{w_j,s}$ are independent for any $i \neq j$. By this assumption, Equation (1) can be rewritten as:

$$\begin{aligned} P(R_{t,s} = 1|t, s) =& 1 - P(\forall w_i \in t, R_{w_i,s} = 0) \\ =& 1 - \prod_{i=1}^{m} (1 - P(R_{w_i,s} = 1)) \end{aligned} \quad (2)$$

As a result, to compute $P(R_{t,s} = 1|t, s)$ it suffices to estimate $P(R_{w_i,s} = 1)$ $(1 \leq i \leq m)$. A simple term matching method does not work here. A sentence s may provide some information about w_i even if w_i does not occur in s. One observation is that while sentences where w_i appears are likely to provide some information about w_i, other sentences closely related to such sentences are also likely to provide information about w_i. We appeal to manifold ranking algorithm [11], where sentences that w_i appears in are chosen as seeds (setting relevance scores to 1). Relevance scores are then propagated among the sentences. Given a topic word w_k, the algorithm that computes the relevance probabilities of the sentences is given as follows. Let $\mathbf{W}_{n\times n}$ be the similarity matrix, where $\mathbf{W}_{ij} (i \neq j)$ is the similarity between sentences s_i and s_j, computed by cosine similarity, and $\mathbf{W}_{ii} = 0$. Normalize $\mathbf{W}$ as $\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is a diagonal matrix with (i,i)-element equal to the sum of the i-th row of $\mathbf{W}$. The manifold algorithm iterates the following equation to get the scores $\mathbf{f}^{w_k} = [f_1^{w_k}, ..., f_n^{w_k}]$ of the sentences.

$$\mathbf{f}^{w_k}(t + 1) = \alpha \mathbf{S} \mathbf{f}^{w_k}(t) + (1 - \alpha)\mathbf{y}^{w_k} \quad (3)$$

where $0 < \alpha < 1$ and $\mathbf{y}^{w_k} = [y_1^{w_k}, ..., y_n^{w_k}]^T$ is set by

$$y_i^{w_k} = \begin{cases} 1 & \text{if } s_i \text{ contains word } w_k \\ 0 & \text{otherwise} \end{cases} , 1 \leq i \leq n \quad (4)$$

Without loss of generality, we initialize $\mathbf{f}^{w_k}(1)$ $(t = 1)$ to $\mathbf{y}^{w_k}$. According to [11], the sequence $\mathbf{f}^{w_k}(t)$ converges to $\mathbf{f}^{w_k *} = (1-\alpha)(\mathbf{I}-\alpha\mathbf{S})^{-1}\mathbf{y}^{w_k}$. Then the relevance probability of s_i to w_k is computed by

$$P(R_{w_k,s_i} = 1) = f_i^{w_k} / (\max_j f_j^{w_k} + \delta) \quad (5)$$

where $\delta > 0$ is a smoothing parameter. For simplicity, $P(R_{w_k,s_i} = 1)$ is abbreviated as p_{ki} from this point on.

By inserting Equation (5) into Equation (2), we obtain the topic relevance probabilities for all sentences in D. These sentences are then ranked in the order of their topic relevance probabilities. The K top ranked sentences might be selected to form a topic-focused summary of D if we did not consider novelty, coverage and balance of the summary.

3.2 Topic-Aware Novelty

Most previous ranking-based approaches achieve novelty by penalizing sentences using an MMR [1] like method. The limitation of this method is that it penalizes each word equally, whether the word is related to a topic or not. In this section, we present an approach that implements topic-aware novelty by extending Equation (2).

Suppose we have already had $k - 1$ sentences $s_1, ..., s_{k-1}$ in a summary. These sentences have covered part of the topic $t = \{w_1, ..., w_m\}$. We want to add a new sentence s_k to the summary such that it covers the most possible of the remaining part of t and the least possible of the part already covered by $s_1, ..., s_{k-1}$. We do this by modifying the topic relevance probabilities of sentences, and use $P(R_k = 1|t, s_1, ..., s_k)$ to denote the relevance probability of sentence s_k to topic t given $s_1, ..., s_{k-1}$. We define

$$\begin{aligned} & P(R_{t,s_k} = 1|t, s_1, ..., s_k) \\ =& P(\exists w_i \in t, R_{w_i,s_1} = 0 \wedge \ldots \wedge R_{w_i,s_{k-1}} = 0 \wedge R_{w_i,s_k} = 1) \end{aligned} \quad (6)$$

This definition states that given sentences $s_1, ..., s_{k-1}$, a sentence s_k is relevant to a topic t if s_k covers at least one topic word w_i that is not covered by any s_j $(1 \leq j \leq k - 1)$. Here we assume that R_{w_k,s_i} is independent of R_{w_t,s_j} for any $k \neq t$ or $i \neq j$. By the same way for Equation (2), expanding Equation (6) leads to

$$\begin{aligned} & P(R_{t,s_k} = 1|t, s_1, ..., s_k) \\ =& 1 - P(\forall w_i \in t, \neg(R_{w_i,s_1} = 0 \wedge \ldots \wedge R_{w_i,s_{k-1}} = 0 \\ & \wedge R_{w_i,s_k} = 1)) \\ =& 1 - \prod_{i=1}^{m} \left(1 - p_{ik} \cdot \prod_{j=1}^{k-1} (1 - p_{ij})\right) \end{aligned} \quad (7)$$

We take the logarithm of $P(R_{t,s_k} = 1|t, s_1, ..., s_k)$ as the *novelty gain* of sentence s_k given $s_1, ..., s_{k-1}$:

$$G_N(s_k) = \log P(R_{t,s_k} = 1|t, s_1, ..., s_k) \quad (8)$$

Obviously, the higher the novelty gain of s_k is, the more relevant to the topic s_k is.

We may construct summaries from documents by selecting one by one the sentences with the highest novelty gains. Such summaries are highly relevant to the given topic and at the same time satisfy topic-aware novelty.

3.3 Topic-Aware Coverage and Balance

Similar to [6, 10], we cluster sentences for coverage consideration. To achieve topic-aware coverage and balance, we first remove all sentences that are not closely relevant to the topic. Since sentences irrelevant to the topic may act as noises in the clustering process, sentence pruning is an important step towards topic-aware coverage and balance. Specifically, we compute sentence relevance probabilities using equation (2) for all sentences. Sentences whose relevance probabilities are lower than a threshold are discarded. After that, we cluster sentences. Then we assign to each sentence a sentence selection preference according to cluster distribution and select sentences based on their selection preferences to achieve coverage and balance. The selection preference function is defined as follows.

Suppose we have built M clusters $(C_1, ..., C_M)$ from the remaining sentences by applying a clustering algorithm such as Kmeans. Denote by $(r_1, ..., r_M)$ the proportions of the clusters, where $r_i = |C_i| / \sum_j |C_j|, 1 \leq i \leq M$, and $|C_i|$ is the number of sentences in C_i. We view r_i $(1 \leq i \leq M)$ as the *expected proportions* of important topic-relevant aspects in each summary. Suppose we have already selected $k-1$ sentences. Let $r_i^{k-1} = N_i^{k-1}/(k-1), 1 \leq i \leq M$ be the distribution of the $k-1$ sentences in the M clusters, where N_i^{k-1} is the number of sentences occurring in $s_1, ..., s_{k-1}$ that belong to cluster C_i. $r_i^{k-1}, 1 \leq i \leq M$ are the actual proportions of important topic-relevant aspects. We want r_i^{k-1} to be as close to r_i as possible to meet coverage and balance requirement. Thus, we score each cluster by

$$score(C_i|s_1, ..., s_{k-1}) = f(r_i - r_i^{k-1})$$

where $f(x) > 0$ and is a strictly monotonically increasing function of x. We take $f(x) = exp(x)$ in this paper. We then define the *selection preference* for cluster C_i as

$$\delta(C_i|s_1, ..., s_{k-1}) = \frac{score(C_i|s_1, ..., s_{k-1})}{\sum_j score(C_j|s_1, ..., s_{k-1})} \quad (9)$$

The above preference shows the degree of suitability under coverage and balance requirements to select the next sentence from C_i, given that $s_1, ..., s_{k-1}$ have already been selected. For each sentence s_k in cluster C_i, we take the logarithm of $\delta(C_i|s_1, ..., s_{k-1})$ as its *coverage and balance gain*:

$$G_{CB}(s_k) = \log \delta(C_i|s_1, ..., s_{k-1}), s_k \in C_i \quad (10)$$

The higher the coverage and balance gain is, the more likely s_k is to cover aspects of the topic different from $s_1, ..., s_{k-1}$.

3.4 A Unified Framework

Suppose we have already selected the first $k-1$ sentences $s_1, ..., s_{k-1}$ for a summary. By combining the novelty gain (Equation (8)) and the coverage and balance gain (Equation (10)), we obtain the following *gain* for selecting the next sentence s_k given $s_1, ..., s_{k-1}$:

$$G(s_k) = G_N(s_k) + \lambda G_{CB}(s_k) \quad (11)$$

where λ is a parameter used to tune the tradeoff between the novelty gain and the coverage and balance gain. Again, the higher the gain of s_k is, the better s_k summarizes the documents w.r.t. the topic. Therefore, to produce the summary, we select sentences that maximize the gain given in Equation (11) one by one.

4. EXPERIMENTS

4.1 Data Sets and Evaluation Metrics

We use the popular topic-focused summarization benchmark data sets DUC 2005, 2006 and 2007[2] for our experiments. The automatic summarizer is expected to extract from each document collection a summary that does not exceed 250 words. We use ROUGE 1.5.5[3] package for evaluation, which is officially adopted by DUC for evaluating automatic generated summaries. Parameter setting for ROUGE is the same as the official parameter setting[4] of DUC. For each document, we use the OpenNLP[5] tool to detect and tokenize sentences. A list of 707 words is used to filter stop words. The remaining words are stemmed by Snowball[6].

4.2 Experimental Results

4.2.1 Overall Performance Comparison

In our experiments we set the parameters empirically. We set α in Equation (3) to 0.95 and λ in Equation (11) to 8. As for the irrelevant sentence pruning, we keep one sixth of the total sentences for each document collection. Cluster number is set to 0.4 times the number of sentences left. We employ two popular clustering algorithms: Kmeans and agglomerative hierarchical clustering. For Kmeans, seeds are randomly chosen, algorithm 2 is run five times, and average performance is collected.

We denote our algorithm by NCBsum, where 'N', 'C' and 'B' stand for novelty, coverage and balance respectively. For simplicity, by "NCBsum-A" we refer to our algorithm where agglomerative hierarchical clustering is applied to cluster sentences, and by "NCBsum-K" we refer to our algorithm where Kmeans is applied.

We compare our algorithm with the following algorithms. (1) Nsum: a method that uses novelty gain (Equation 8) to select sentences. (2) Csum: a method that considers topic-aware coverage. It prunes irrelevant sentences according to Equation (2), clusters the remaining sentences and selects the most relevant sentences from different clusters. (3) Coverage: a baseline clustering-based method. It clusters the sentences and select the most relevant sentences from different clusters. (4) Manifold: the manifold ranking algorithm proposed by Wan et al. [8] without considering novelty, coverage and balance. (5) Auto avg: the average scores of the participating systems in DUC 2005, 2006 and 2007 respectively. (6) Random: a baseline algorithm that produces summaries through random sentence selection.

Tables 1, 2 and 3 show the experimental results. Scores in bold are the best scores. Scores with * mean the performance improvement over the Manifold Ranking algorithm is significant with confidence level 95%. Clearly, our algorithm outperforms the comparative algorithms on all the f-measures. Our algorithm NCBsum-A ranks 1st, 2nd and 4th among all the participating systems in DUC 2005, 2006 and 2007 respectively.

[2] Refer to http://www-nlpir.nist.gov/projects/duc/data.html for a detailed description of the datasets.

[3] http://www.isi.edu/publications/licensed-sw/BE/index.html

[4] -n 4 -w 1.2 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a -d

[5] http://opennlp.sourceforge.net/

[6] http://snowball.tartarus.org/index.php

	ROUGE-1	ROUGE-2	ROUGE-W
NCBsum-A	**0.38868***	**0.07878***	**0.13424***
NCBsum-K	0.3857*	0.07678	0.13282*
Nsum	0.38169	0.07507	0.13057
Csum	0.3826	0.07657	0.13224
Coverage	0.37328	0.071	0.1289
Manifold	0.37493	0.0741	0.12916
Auto Avg	0.34347	0.06024	0.11675
Random	0.30994	0.03892	0.10616

Table 1: F-measure comparison on DUC05

	ROUGE-1	ROUGE-2	ROUGE-W
NCBsum-A	**0.40869***	**0.08981***	**0.14074***
NCBsum-K	0.40515*	0.08698*	0.13972*
Nsum	0.4026	0.08501	0.13787
Csum	0.40354	0.08603	0.13876
Coverage	0.39811	0.08383	0.13705
Manifold	0.38813	0.08168	0.13396
Auto Avg	0.37959	0.07543	0.13001
Random	0.34421	0.0513375	0.1177875

Table 2: F-measure comparison on DUC06

	ROUGE-1	ROUGE-2	ROUGE-W
NCBsum-A	**0.4289***	**0.11139***	**0.1478***
NCBsum-K	0.42244*	0.10931*	0.1464*
Nsum	0.422	0.10752	0.14499
Csum	0.4246	0.10861	0.14739
Coverage	0.42242	0.10116	0.14531
Manifold	0.40056	0.10106	0.13859
Auto Avg	0.40048	0.09544	0.13728
Random	0.36193	0.06326	0.12324

Table 3: F-measure comparison on DUC07

We would like to compare our algorithm with Qs-MRC [9], which to the best of our knowledge, is the only algorithm that considers topic-aware novelty in summarization. We notice that only ROUGE recall scores on DUC 2005 are provided in that paper. Therefore we compare recall of our algorithm with that of Qs-MRC, which are shown in table 4. Systems 15 and 17 are the two best participating systems in DUC 2005. Obviously, NBCsum-A outperforms Qs-MRC in all recall measures.

	ROUGE-1	ROUGE-2	ROUGE-SU
NCBsum-A	**0.39092**	**0.07924**	**0.13794**
Qs-MRC	0.3868	0.0779	0.1366
System 15	0.3751	0.0725	0.1316
System 17	0.3697	0.0717	0.1297

Table 4: Recall-measure comparison on DUC05

4.2.2 Evaluation on Parameter λ

λ (Equation (11)) is a parameter used to tune the trade-off between the topic relevance probability with the topic-aware novelty and topic-aware coverage and balance. We did experiments with different λ values to see how it affects the quality of summaries. Figures 1 shows the experimental results with λ varying from 0 to 512. On the whole, performance of the algorithm first improves and then degenerates with the increase of λ.

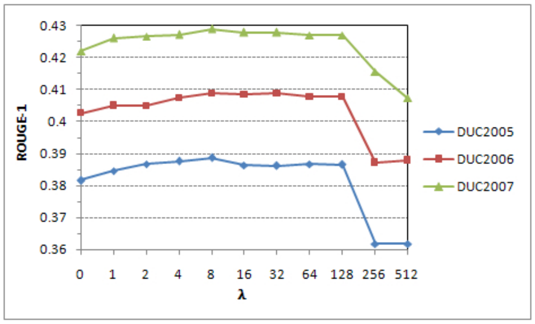

Figure 1: ROUGE-1 vs. λ

5. ACKNOWLEDGMENTS

This work is supported in part by NSFC grants 60970045 and 60721061, and by the National High-tech R&D Program (863 Program).

6. REFERENCES

[1] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proc. of SIGIR '98*, 1998.

[2] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proc. of SIGIR '08*, 2008.

[3] G. Erkan and D. R. Radev. Lexpagerank: Prestige in multi-document text summarization. In *Proc. of EMNLP '04*, 2004.

[4] L. Li, K. Zhou, G.-R. Xue, H. Zha, and Y. Yu. Enhancing diversity, coverage and balance for summarization through structure learning. In *Proc. of WWW '09*, 2009.

[5] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In *Proc. of EMNLP '04*, 2004.

[6] T. Nomoto and Y. Matsumoto. A new approach to unsupervised text summarization. In *Proc. of SIGIR '01*, 2001.

[7] X. Wan and J. Yang. Multi-document summarization using cluster-based link analysis. In *Proc. of SIGIR '08*, 2008.

[8] X. Wan, J. Yang, and J. Xiao. Manifold-ranking based topic-focused multi-document summarization. In *Proc. of IJCAI'07*, 2007.

[9] F. Wei, W. Li, Q. Lu, and Y. He. Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In *Proc. of SIGIR '08*, 2008.

[10] H. Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proc. of SIGIR '02*, 2002.

[11] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. In *Proc. of NIPS '03*, 2003.